DOCUMENT RESUME

ED 474 143 HE 035 676

AUTHOR	Luan, Jing
TITLE	Data Mining and Knowledge Management in Higher Education - Potential Applications.
PUB DATE	2002-06-00
NOTE	19p.; Paper presented at the Annual Forum for the Association for Institutional Research (42nd, Toronto, Ontario, Canada, June 2-5, 2002).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Cluster Analysis; Data Analysis; *Decision Support Systems; *Educational Planning; *Prediction; *Reentry Students
IDENTIFIERS	*Data Mining; *Knowledge Management

ABSTRACT

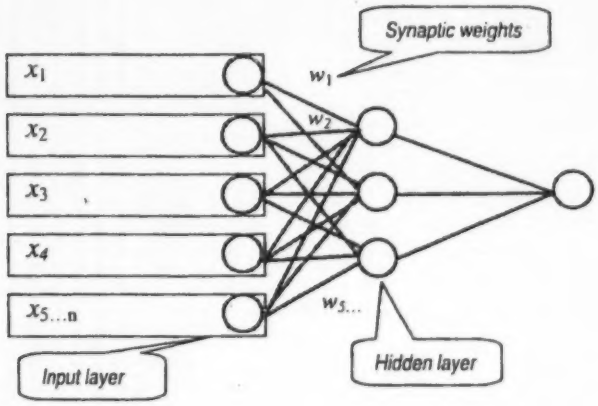
This paper introduces a new decision support tool, data mining, in the context of knowledge management. The most striking features of data mining techniques are clustering and prediction. The clustering aspect of data mining offers comprehensive characteristics analysis of students, while the predicting function estimates the likelihood for a variety of outcomes, such as transferability, persistence, retention, and success in classes. Compared to traditional analytical studies that are often hindsight and aggregate in nature, data mining is forward looking and is oriented to individual students. A real life project presents the work of data mining in predicting the possibility of return to school for every student currently enrolled at a community college in Silicon Valley. The project applies neural network techniques and two rule induction algorithms, C&RT and C5.0, to choose the best prediction followed by a clustering analysis using TwoStep. The list of students who are predicted as less likely to return to school by data mining is then turned over to faculty and management for direct and indirect intervention. The paper also discusses potential applications of data mining in higher education. The benefits of data mining are its ability to gain deeper understanding of the patterns previously unseen using current available reporting capacities. In addition, prediction from data mining allows the college an opportunity to act before a student drops out or to plan for resource allocation with confidence gained from knowing how many students will transfer or take a particular course. An appendix discusses Cross Industry Standard Procedures for Data Mining. (Contains 4 tables and 10 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 474 143

DATA MINING AND KNOWLEDGE MANAGEMENT IN HIGHER EDUCATION

-POTENTIAL APPLICATIONS



PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Vura

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

- What is data mining?
- Where does data mining fit in the context of Knowledge Management?
- What is a backpropogating neural net, entropy, decision tree?
- How to transfer data mining techniques developed for the corporate world to higher education?
- Which data mining model is the best for predicting a student's chance of persistence? Retention? Success?
- Who are the ones that are transferring or dropping out?
- Read on…

BEST COPY AVAILABLE

Jing Luan, Ph.D., ITM

HE035076

EXECUTIVE SUMMARY

This paper introduces a brand new and powerful decision support tool, data mining, in the context of knowledge management. Among other things, the most striking features of data mining techniques are clustering and prediction. The clustering aspect of data mining offers comprehensive characteristics analysis of students, while the predicting function estimates the likelihood for a variety of outcomes of them, such as transferability, persistence, retention and success in classes. Compared to traditional analytical studies that are often hindsight and aggregate, data mining is forward looking and is oriented to individual students.

A real life project presents the work of data mining in predicting the possibility of returning to school for every student currently enrolled at a community college in Silicon Valley. The project applies neural network, C&RT and C5.0 to choose the best prediction followed by a clustering analysis using TwoStep. The list of students who are predicted as less likely to return to school by data mining is then turned over to faculty and management for direct or indirect intervention.

The paper also discusses potential applications of data mining in higher education. The benefits of data mining are its ability to gain deeper understanding of the patterns previously unseen using current available reporting capabilities. Further, prediction from data mining allows the college an opportunity to act before a student drops out or to plan for resource allocation with confidence gained from knowing how many students will transfer or take a particular course.

Copyright . . 2002.

Data Mining and Knowledge Management in Higher Education -Potential Applications

Introduction

An item soon to be high on the agenda for researchers and administrators in higher education is the adoption of data mining. Higher education will find larger and wider applications for data mining than its counterpart in the business sector, because higher education institutions carry three (3) duties that are data mining intensive: scientific research that relates to the creation of knowledge, teaching that concerns with the transmission of knowledge, and institutional research that pertains to the use of knowledge for decision making. All the above tasks are well within the boundaries of Knowledge Management, which drives the need for better and faster decision making tools and methods. Scientific research in chemistry, physics and the like uses data mining for pattern recognition and cluster detection. Using data bits collected from precision measuring tools for this type of research, no matter how large the dataset, is not nearly as challenging as the other type of scientific research: social sciences, psychology, where researchers encounter missing data, fuzzy measurements and unavailable attributes. Chief among them is Institutional Research. Other entities that have suddenly found the elevated applications of data mining, such as the FBI for behavior pattern identification, have indeed opened an exciting chapter for the wide use of data mining in general and higher education in particular. This white paper is written for researchers with intermediate data mining knowledge in mind, but beginning users will definitely find it informative. Readers in businesses outside higher education will also find it useful.

Knowledge Management Driving Data Mining

Several authors have written about the factors behind the dawn of data mining. For instance, Therling (1995) identified three (3) reasons: The ease of data collection and storage, the computing power of modern processors, and the need

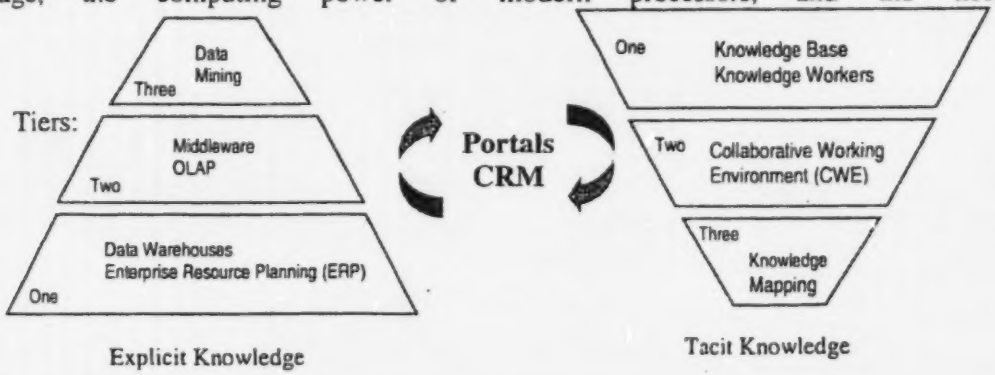


Figure One. Tiered Knowledge Management Model (TKMM).

for fast and real time data mining. Yet, one important reason absent from these is the growing interest in Knowledge Management. Knowledge, a focal point of ontology or

epistemology, is the product of moving from data to information and finally to knowledge. The following model, Tiered Knowledge Management Model (TKMM) developed by Jing Luan (2000, 2002) illustrates the dichotomous nature of modern knowledge management framework for higher education research professionals.

The components of Knowledge Management are Explicit (documented, measurable) and Tacit (subjective, qualitative). Documented, measurable explicit knowledge is most familiar and available to us, as it exists mostly in databases and other similar medium. While tacit knowledge, an entity of feelings, personalities and aptitudes (Crowly, 2000; Davenport & Prusak, 1998) is crucially important, but it is hard to quantify and is not the focus of this thesis. Customer Relationship Management (CRM) and portals are the mechanisms that make both types of knowledge work together for a business purpose. All three components in CRM, operational, analytical and collaborative, are key users of data mining.

On the Explicit side, data mining reflects the highest level of knowledge attainment that requires skills in data domain (Tier One), data querying and presentation (Tier Two) and artificial intelligence/machine learning (Tier Three). Data mining occupies the top tier and is dependant on the lower tiers. The following chart is a topology of the Explicit Knowledge of TKMM that illustrative in detail the relationships among three tiers and the software programs for each:

Figure Two. Topography of Tiered Knowledge Management Model (TKMM) for explicit knowledge

Data Mining Definitions

The definition from Gartner Group seems to be most comprehensive, as they define data mining as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern

recognition technologies as well as statistical and mathematical techniques." The author has refined the notion of data mining for higher education to be a process of uncovering hidden trends and patterns that lend them to predicative modeling using a combination of explicit knowledge base, sophisticated analytical skills and academic domain knowledge. It is producing new observations from existing observations. Or, as explained by Rubenking (2001), "data mining is the process of automatically extracting useful information and relationships from immense quantities of data. In its purest form, data mining doesn't involve looking for specific information. Rather than starting from a question or a hypothesis, data mining simply finds patterns that are already present in the data."

Applications of Data Mining in Higher Education

What are the transferable techniques in data mining that are readily applicable in higher education? In fact, there are many. Algorithms are similar in concept to stored procedures in object related programming in that they are universally applicable. Almost all algorithms or models currently used in the business sector are directly usable for research in higher education, especially in institutional research.

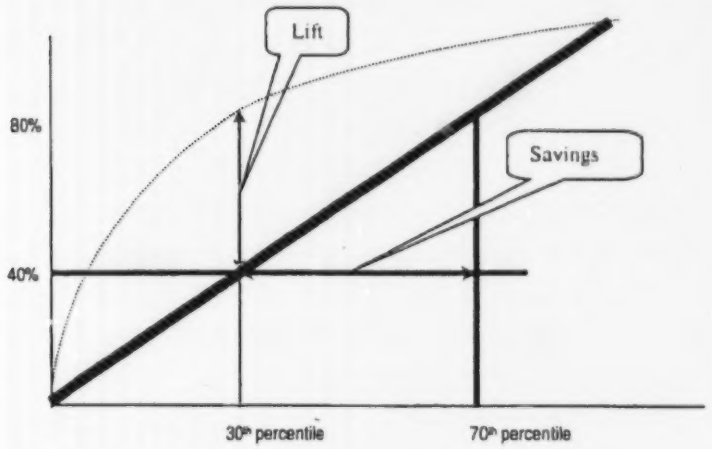
The table below is a typical translation of the type of data mining questions applicable for use in higher education.

Table One: Data mining questions in the business sector and their counterpart in the higher education sector:

Bottom-line Questions in the Business World	Counter-part Questions in Higher Education
Who are my most profitable customers?	Who are the students taking most credit hours?
Who are my repeat website visitors?	Who are the ones likely to return for more classes?
Who are my loyal customers?	Who are the persisters at our university/college?
Who is likely to increase his/her purchases?	Which alumnus is likely to donate/pledge more?
What clients are likely to defect to my rivals?	What type of courses can we offer to attract more students?

One way of understanding what data mining can usually bring to the filed of higher education is to use Gain Charts. The chart below is a Lift chart, one of the most frequently used tools to examine the effectiveness and benefits of conducting data mining. For example, if certain alumni contribution patterns are identified for an institution with 50,000 alums who each contributes $100 a year, just one percent increase can mean $50,000. The amount will greatly increase if generous contributors are better identified. This is the so- called One Percent Doctrine (Luan, 2000). To further illustrate the point of gains in data mining, the following gain chart (one type of lift charts) uses a hypothetical alumni pledge mailing.

The lift is the difference of using or not using data mining techniques. The equation is Lift = $P(\text{class}_t|\text{sample})/P(\text{class}_t|\text{population})$, where probability of P is a function of the purposely selected biased sample from the general population.



Gain Chart showing the benefit of Data Mining:

In a hypothetical mass mailing for alumni pledge, set the dotted line to be the optimal return rate (alumni sending in contributions) as predicated by data mining and the straight 45-degree line the result if the entire population received the mailing. If data mining were used, when the mailing reached about 30th percentile of the population predicted to be "responsive", 80% of them would have responded vs. 40% without data mining. If every percentage point = $2,500, savings =(70% * $2,500) – (30% * $2,500) = $175,000 - $75,000 = $100,000). Without data mining, it would cost $100,000 more to reach all 80%.

Figure Three: Lift chart showing the benefits of data mining for alumni pledge.

POTENTIAL APPLICATIONS OF DATA MINING IN HIGH EDUCATION

Everything is data, but data isn't everything. Therefore, in a data rich society in which we increasingly find ourselves in, there are a myriad of opportunities for data mining. There are several ways to examine the potential applications of data mining. One is to start with the functions of the algorithms to reason what they can be utilized for. Another is to examine the areas of an institution where data are rich, but mining activities are scarce. And another is to examine the different functions of a university or college to identify the needs that can translate themselves into data mining projects.

Some of the most likely places where data miners (institutional researchers who wear this hat) may initiate data mining projects are:

- Alumni
- Institutional Effectiveness
- Marketing
- Enrollment Management

All universities are committed to alumni management, but not all of them are using the best tools. Data mining is this area can potentially help identify those who are most likely to donate or participate in alumni related activities. For big donors, data miners can use the method of treating outliers.

Institutional Effectiveness, synonymous with assessment of learning, but perhaps larger, is having sweeping impact on higher education. How do students learn best? What courses are often taken together? What learning experiences are most contributive to

overall learning outcomes? These intriguing questions are good candidates for data mining. National organizations are moving quickly into this field. Nationally well known projects, such as NSSE (National Survey of Student Engagement) and its sister CCSSE (Community College Survey of Student Engagement), CSEQ (College Student Experience Questionnaire) have accumulated enough data that quickly resemble situations in the private industry where huge amount of consumer feedback becomes fertile grounds for clustering and predictive modeling.

Marketing is in a state of revival in higher education. As modalities of delivering learning have changed from "only game in town" to "anywhere, anytime", institutions began to find themselves locked in a semi-fierce to fierce competition. Who else has the college not reached? Who may be interested in receiving more information in a particular program area? Marketing, powered by data mining, may provide just the right amount of "lift" to bring in more enrollments.

As colleges and universities reexamine themselves, a painful experience by all means, enrollment management will be re-engineered and re-invented. Enrollment management was met with a variety of realistic challenges in the late 80s because it could not find the speed and tools to quickly identify the prospective student, to pin-pointed the time when a student may drop out and to meet the needs for services as quickly as students would desire. With data mining, web technology, and future Learner Relationship Management (LRM), enrollment management is once again poised to becoming a key player in higher education.

A Case Study of Predicting and Clustering Persisters and Non-persisters

In both four year and two-year institutions, as a matter of choice, persistence has become an indicator of academic performance and enrollment management. In a broad sense, more students who persist means better academic programs and higher revenue. Therefore, one of the tasks important to an institution's management and its faculty is to identify those students who are less likely to return from semester to semester. Having these students identified soon enough for the institution to tailor its interventions and marketing strategies will greatly enhance the persistence rate. Various types of studies exist to analyze cohort based persistence patterns and rates. However, it remains an illuminating task to be able to predict a currently enrolled student's likelihood of returning to school the next term. The ability to accomplish this will most certainly give the institution a jumpstart.

With the experience in building and maintaining one of the first higher education college based data warehouses dating back to the early 90s, the author chose a typical community college data warehouse on the west coast for this project. The project (notice it is not called a "study") chose a recent semester (fall 2000) as the end term in the dataset. Although many fields within a dataset could be either output or input (bi-directional) field, the first task was to decide on a primary output field for this data mining task. The action of students who either re-enrolled or did not re-enroll the next term (spring 2001) became the output field with "P" to denote persisters and "NP" non-persisters. Using a

technique developed by Knowledge Discovery Laboratory, the data were split evenly across the fields into a training set and a test set.

Following Steps for Data Mining Data Preparation developed by the author (see appendix) and the principals of CRISP-DM (Cross Industry Standard Process for Data Mining), the author built two datasets (feature sets) with one for testing and the other for validation. Data mining is an iterative process and identifying patterns is even more so. The process was two-fold. First, train the models using the datasets with several different algorithms jointly, called "bagging" to identify one or two models that consistently produced optimal predictions on existing data. The use of several techniques to cross validate particular extrapolatives, including classical statistical techniques, is a recommended approach. Secondly, the project brought in brand new data from the current semester for actual scoring. The actual process to build the dataset took three weeks, as both the time span of the dataset and the sheer number of fields (n=164) with many calculated were pushing the limit of the available software and hardware. Having to spend a disproportionate amount of time on building datasets is typical of data mining projects – of which a data mining research professional must be mindful.

The following is a partial list of the groups of features (fields) selected for this case study. They are:

- Demographics: Age, Gender, Ethnicity, High School, Zip Codes, Planned Employment Hours, Residence, Education Status at Initial Enrollment
- Total Transfer, Vocational, Basic Skills, Science, Liberal Arts Courses Taken
- Number of Courses Taken by Department
- Total Units Earned, and Grade Points by Course Type

Clementine, an award winning software by SPSS Inc. enjoys the reputation for being the easiest to deploy and one that has incorporated major data mining models in its tool chest. The project chose Clementine as the data mining software precisely for this area of strengths. Experiences led the author to use Neural Networks (NN) and two rule induction algorithms, C5.0 and C&RT, to compare models and to compliment the scoring. The Type node in Clementine automatically instantiated the field type appropriately. Since half of the records were persisters and half were not, there was no need to balance the dataset within itself.

Artificial neural networks (ANN) are synonymous with data mining, therefore it is a good starting place for using an ANN architecture to illustrate the basic data mining model as well as the various relationships in the model. Figure Three is a typical backpropagating neural network, also called Multi-layer Perceptron (MLP).

Expressed in a mathematical formula, the model is

$$o_j = f\left(\sum_{i=1}^{n} o_i w_{ji}\right), \text{ where}$$

$$f'(x) = -(1+e^{-x})^{-2} e^{-x}(-1)$$

where o is the outcome, x_i is the input vector, w_i is the weight, which is set at random by the model upon first feed. Some of the input fields (vectors) are in groups, while others are single fields. The output contains two outcomes of P and NP. The hidden layer performs the summation and constantly adjusts the weights until it reaches an optimal threshold, which then determines the outcome for a record (Garson, 1998). In this case, the features of Number of Terms Enrolled, GPA, and so on are inputs. The model adjusts its weights repeatedly until it deems optimal or is stopped by the researcher. The actual model used by the project contained two hidden layers, which is explained later.

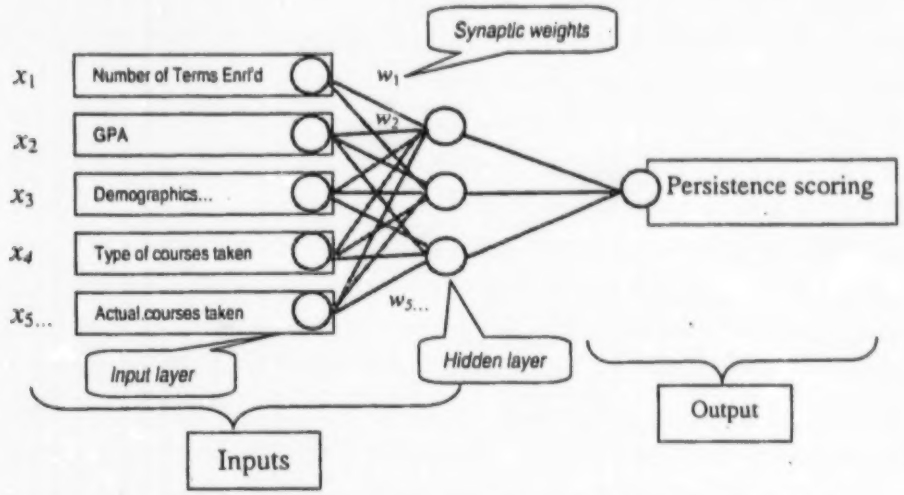


Figure Three: Diagram of the Backpropagating Neural Network for Persistence Modeling.

The author first trained the neural network, called "Train Net Node" in Clementine. This node usually produces a list of fields based on their Relative Importance as Inputs, which can be helpful for a number of purposes. Below, the node shows such a list of "Relative Importance as Inputs" abridged by the author for brevity as follows:

Educational Goal	(.39910)
Probation Status	(.34453)
Enrollment Status	(.32316)
Hours of Work Per Week	(.27844)
The # of Terms Enrolled	(.26942)
Residence Status	(.20012)
High School Attended	(.19680)
Art History Course Taken	(.19488)
Allied Health Course Taken	(.18625)
Applied Living Arts Course	(.18511)
...	
n=145.	

Theoretically the values would range from .0 to 1.0 with 1.0 being extremely important. In practice, the fields rarely go above .32 with the obvious exception of Educational Goal

BEST COPY AVAILABLE

in this case. A couple of points worth noting are that the investigator should pay attention to every field even the one listed at the bottom, as data mining is both a task to identify the averages, rule of thumbs and a task to use outliers for a number of reasons, such as fraud detection.

The Train Net Node resulted in an accuracy of 61.3%. It contained 190 input neurons, 20 neurons in hidden layer one, 15 neurons in hidden layer two, and 1 output neuron (for a dichotomous output). The difference between the total fields entered and the number of input neurons lies in the number of values in set fields.

With this in mind, let's examine two other popular algorithms used by the project for prediction. One is C5.0 and the other C&RT (Classification and Regression Trees). C&RT handles binary splits the best, while multiple splits are best taken by C5.0. These two algorithms differ in the criterion used to drive the splitting. C5.0 relies on measures in the realm of the Information Theorem and C&RT utilizes the Gini coefficient (SPSS, 2000). Rule induction is fundamentally a task of reducing the uncertainty (entropy) by assigning data into partitions within the feature space based on information-theoretic approaches (Eijkel, 1999). The mathematical formula on discerning uncertainty is expressed as measurements in bits:

$$H(N) = \sum_{n=1}^{n} -P(n) \log_2 P(n)$$

where H (N) is the uncertainty defined as discrete information and P(n) is the probability that $\xi = n$. As uncertainty reduces, bits are reduced.

Even though Occam's Razor dictates that a simpler explanation is more likely to capture the essence of a data mining problem, it was not the case in this project due to the nature of the datasets and the specific scoring task at hand. Due to the "loosey" nature of data, hardly any social science data mining research projects can be considered robust without painstaking fine tuning or using several algorithms all at once. Many researchers have discussed the pruning tasks associated with using these heuristics based algorithms. Sometimes it is a judgment call, and sometimes it is sheer luck. For this project, the first few runs did not produce acceptable results that were "actionable". The prediction rate was never more than 65% accurate for either P or NP. An examination of the data was necessary.

There were two characteristics in the datasets that might have contributed to the level of difficulty in securing the highest prediction accuracy. First, the dataset contained longitudinal data dating back more than 5 years, which had a number of changes in course naming convention and shifting in teaching methodologies. Secondly, students in community colleges have low "loyalty" and many of them tend to concurrently enroll in two or three colleges or jump from one to the next frequently.

Since the issue at hand was to correctly predict those who were not returning the next semester, a decision was made to weigh the cost of misclassification. The project opted to gain accuracy in predicting non-persisters at the expense of predicting accuracy for persisters. In other words, the decision makers would rather treat some of the persisters as

non-persisters when conducting targeted marketing than the other way around. This decision resulted in a prediction accuracy rate for the training set to be 85.1% for the NP (Non-persisters) and 27.5% for the P (persisters). C5.0's prediction accuracy for the training set was 45.5% for P and 77.1% for NP, while C&RT's prediction accuracy for the training set was 57.2% for P and 63.1% for NP.

Table Two: Outcomes of Persistence Modeling for Test and Training Sets:

Neural Net on Test Set*

	NP	P
NP	**86.36**	13.7
P	72.4	**27.6**

On Validation Set

	NP	P
NP	**85.1**	15.0
P	72.5	**27.5**

C5.0 on Test Set**

	NP	P
NP	**85.3**	14.7
P	46.9	**53.1**

On Validation Set

	NP	P
NP	**77.1**	22.9
P	54.5	**45.5**

C&RT on Test Set

	NP	P
NP	**43.7**	56.3
P	21.1	**78.9**

On Validation Set

	NP	P
NP	**42.1**	57.9
P	22.7	**77.2**

* Using the Expert Option, the author chose prune and 2 hidden layers, which greatly increased the accuracy for predicting non-persisters at the expense of accuracy for persisters.

** Cost of 2.0 if Non-persisters are classified as persisters at 75 pruning severity.

As shown above, the neural network node has produced the best model for this purpose. However, beyond the list of Relative Importance as Inputs, the neural network node does not provide any further information, cryptic or not. The transparency of a neural network is inherently non-existent. Researchers have long deemed what takes place inside the neural network as "black box" magic. While the project will eventually choose neural net model for scoring, it is necessary to use a rule induction model to list the rules uncovered in order to better understand how the model has worked.

C5.0 can generate both decision trees and rule sets. Rules are easier to understand, as it most closely resembles the task of profiling. There were 13 rules for Persisters and 38 rules for Non-persisters produced by C5.0. As an example, a typical rule for persisters reads:

If # of Courses Taken > 7
    and Probation Status = Good
    and Applied Living Arts Courses Taken = 0
    then P (426, 0.743)

A typical rule for non-persister reads:

If # of Courses Taken > 7

Copyright [illegible], 2002.

and Basic Skills Courses Taken <=1
and Probation Status = Poor
and Applied Health Course Taken <=0
and Art History > 1
then NP (104, 0.585)

Even though the above rules are highly useful in understanding how the model has worked, they do not assist in understanding what is in the dataset. As a matter of fact, researchers need to first visualize and cluster the dataset (unsupervised data mining) before conducting classification and estimation (supervised data mining). Supervised data mining algorithms are very useful when a researcher is very knowledgeable about the data domain and is confident in choosing a particular output field. There are occasions when the researcher must fish around to understand the various permutations and groupings within a hyperspace in order to make the information actionable. These methods of gaining insights into the feature space are called unsupervised data mining algorithms. The following table provides information on tasks for which currently available data mining algorithms can be used.

Table Three: Categorization of data mining algorithms:

TASKS	SUPERVISED	UNSUPERVISED
Classification	Memory Based Reasoning, Genetic Algorithm, C&RT, Link Analysis, C5.0, ANN	Kohonen Nets
Estimation	ANN, C&RT	...
Segmentation/ Association	Market Basket Analysis, Memory Based Reasoning, Link Analysis, Rule Induction	Cluster Detection, K-means, Generalized Rule Induction, APRIORI, Two-Step
Description	Rule Induction, Market Basket Analysis	Spatial Visualization

Kohonen and Kmeans, compared to TwoStep, are a bit cryptic to interpret and somewhat demanding in the types of fields for inputs. The project chose TwoStep Cluster technique for its tolerance of diverse data types and user-friendly groupings. Note, this is not an exercise to establish typologies, in which case, far more manual categorization should have occurred prior to actual modeling. One way of understanding groupings typically involves examining a secondary level of factors associated with the main outcomes of the data mining project. This would mean going beyond persisting and non-persisting, transfer and non-transferred to a level that define when or how the outcome happened, for example, number of terms prior to a student became transfer ready, or number of courses continually taken by a student prior to becoming transfer ready. In brevity, the following clustering analysis represents a general analysis of the entire population to seek major centroids. Since data mining is iterative work, this part of the analysis may occur before predictive modeling is conducted, so that somewhat homogenous populations exist to make the predicted score more precise. A subjective number of clusters of six (6) was set

by the author for the model. For illustrative purposes, two (2) distinctive clusters from the six (6) are as follows:

Cluster 2: (Total 610 records)
Numeric Fields:

- Number of Courses Taken: 17.2 (10.0)
- Units: 25.1 (20.9)
- Degree Applicable Courses Taken 11.8 (8.9)
- Non-Degree Applicable Courses Taken 5.4 (6.5)
- University Transferable Course Taken 5.5 (5.7)
- Non-transferable Courses Taken 9.9 (7.8)
- Pre-collegiate Basic Skills Taken 2.6 (3.1)

Symbolic Fields:

- Hispanics (59.2%)
- Outside County High School (73.3%)
- High School Graduates (40.5%)
- Female (63.4%)
- Non-English Speakers (61.7%)

Cluster 3: (Total: 1,477 records)
Numeric Fields:

- Number of Courses Taken: 8.5 (6.0)
- Units: 15.6 (13.5)
- Degree Applicable Courses Taken 8.2 (5.9)
- Non-Degree Applicable Courses Taken .4 (.9)
- University Transferable Course Taken 4.0 (4.1)
- Non-transferable Courses Taken 2.9 (3.6)
- Pre-collegiate Basic Skills Taken .1 (.4)

Symbolic Fields:

- White (84.8%)
- Post High School Degrees (54.2%)
- Female (64.6%)
- English Speakers (97.9%)

The distinct characters between these two clusters are highly noticeable in both demographics and course taking patterns. An exercise like this is indeed valuable to assist researchers to evaluate the supervised modeling techniques and eventual decisions. Conceivably, upon learning the characteristics of the dataset, the researcher may choose to reduce the less crucial data fields, especially if processing time or storage is a concern. The use of factor analysis/principal components analysis (PCA) to root out the auxiliary features may be desirable at this point.

A Persistent Question From Researchers

Since data mining traces its lineage back to statistics and technology, it is only natural for researchers to inquire about the differences among the three fields. As a matter of fact, this is a question very frequently brought up among researchers. To address this question, it is necessary to base the answer on the three tiers of the knowledge management model. Delmater and Hancock (2001, p192) wrote, "The science underlying predictive modeling is a mixture of mathematics, computer science, and domain expertise." Their point is very important because it refers to the three tiers in TKMM. In the following table, the author has produced a high level view of the comparable features among data mining, statistics and data warehouse based OLAP. The intended purpose of the following crosswalk is for researchers to have a broad idea of how to relate what they are familiar with to what they are learning from data mining.

Table Four: Crosswalk of data mining models/algorithms to Statistics and Data Warehouse Based OLAP.

DATA MINING	STATISTICS	DATA WAREHOUSING/OLAP
Artificial Neural Networks	Regression equations, Chi-square, Structural equations	...
Rule Induction	Principle Components, Discriminant Function, Factor Analysis, Logistic R	...
Kohonen Networks	Cluster Analysis, Probability Density Function	Multi-dimensional cube
Spatial visualization	2-3 dimension charts	2-3 dimension charts
Euclidean Space	Structured equations, Linear and non-linear regression	Sequential files
Classification	Logistical regression	Multi-dimensional cube
Estimation	Regression equations, Chi-square, Structural equations	...
Segmentation	Cluster Analysis, Factor Analysis	Multi-dimensional cube
Prediction accuracy	Statistical significance	Temporal/trend reporting
Outliers detection	Standard deviation, error analysis	Aggregation
Supervised learning	Hypothesis, distributional assumptions, a-priori	...
Unsupervised learning	Descriptive statistics, Cluster Analysis	Temporal/trend reporting
Population/universe	Samples	Fact tables and dimension tables
Feature vectors	Histogram, correlation	Cross-tabs
Feature Extraction	Flat files	Extract, transform, load (ETL)
Machine learning/artificial intelligence	Mathematics	Structured Query Language (SQL)
Attributes, features	Variables, values	Fields, records
Outputs or scoring	Independents	Fields
Population in Current Time Slice	Cohorts, samples	Multi-dimensional cube

As expressed earlier in the chapter, basic and classical statistical knowledge is highly useful to a data mining professional in discerning minute significance in cluster boundaries – an important data mining task. Neural networks are in essence regression models adapted to conduct estimation. In this sense, what was a concern to a researcher, statistical significance, is now a concern of how it translates into accuracy of the prediction or classification. In the final analysis, data mining research questions do not begin with "what *if*", instead, they begin with "what *is*".

Tools for data mining are constantly emerging and there are perhaps as many vendors of data mining software as data mining techniques. The most familiar algorithms in the data mining community are ANN, C&RT, C5.0, CHAID (Chi-squared Automatic Induction), K-means, Nearest Neighbor, MBR (Memory Based Reasoning), and Automatic Cluster Detection. Less known or specialized algorithms are Link Analysis, MBA (Market Basket Analysis, Genetic Algorithms, and Fuzzy Logic.

Conclusion

Chief among a host of recent technology innovations, data mining is making sweeping changes to the entire makeup of our skills and comfort zones in information analysis in higher education. Not only does it introduce a surfeit of new concepts, methods and phrases, it also departs from the well-established traditional hypothesis based statistical techniques. Data mining is a new type of exploratory and predictive data analysis, which is long overdue. It works well in delineating systematic relations between variables when there are no (or not complete) a priori expectations as to the nature of those relations. Data mining has tremendous applications in higher education institutional research alone. They range from marketing, alumni fund raising, to survival analysis, persistence and many others.

Individuals who are data domain experts, learned in classical statistics, and possess the abilities to implement IT systems will find it a smooth sail when transitioning from current decision support environment to knowledge management driven data mining environment. Further, interpersonal communication skills are a plus when needed to convince upper management to adopt data mining. One notion often found useful to institutions that are beginning to consider data mining is the one-percent doctrine. For an institution with an enrollment of 15,000 students with each student paying tuition as low as $5,000, the tuition revenue is $75 million. If data mining is able to increase enrollment by 1%, this will represent $750,000. The same idea can be applied to alumni pledge, graduation, and others. If degree attainment is of the utmost interest to university administrators, the persistence modeling introduced in this paper is aimed at helping move the students along the pipeline with fewer dropouts as possible. This, unquestionably, translates into increased efficiency, higher graduation rates, and lower cost to society as a whole.

APPENDIX

CRISP-DM and Its Application

CRISP-DM stands for Cross Industry Standard Procedures for Data Mining. Why should there be standards? Briefly, a data miner needs to know more than just statistics or database technologies to perform competently. Data mining requires knowledge in the following areas: database/data warehousing, statistics, domain expertise and business processes. As the demand for data mining is growing exponentially and more algorithms are created and put to use, there must be good practices that everyone can follow, much like ISO9000 for manufacturing, GAAP for accounting, and accreditation standards for measuring learning in education. Data mining is not just a study or a report. It is a dynamic business operation and mission critical part of customer relationship management (CRM). It is a project that is highly structured. As with any project, there are phases, stages, steps, assessment and planning. CRISP-DM addresses each of the areas at specific junctures with detailed dos and don'ts.

There are six phases currently recognized in any given data mining project: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. For instance, in the first phase of Business Understanding, the data miner needs to determine the business objective (what are we trying to achieve?), assess the situation (what is available? who does what?). In the Modeling phase that people have read about the most, there are what we discussed in this paper steps involving selection of modeling techniques, conducting iterative testing on data, and evaluating the models. For each step, there are tasks underneath. The concept of describing data mining as throwing everything in the kitchen sink to see what comes out on the other end of the mining stream is crude and is beginning to mislead people. A data miner may end up spending 80% of the time in Data Preparation phase, which may be mentioned in only one sentence buried deep in a report. That should not diminish the importance of the Data Preparation phase.

Knowledge Discovery in Database (KDD) holds the key to the success of data mining. KDD means that the stabilized data mining algorithms produced by the data miner are superimposed over the live database, sometimes directly over the college's ERP (Enterprise Resource Planning) databases, such as Peoplesoft, SAP, Datatel, in order to "score" live data. Without the guidance of CRISP-DM, KDD will encounter many stumbling blocks.

STEPS FOR DATA MINING PREPARATION
(Based on CRISP-DM)

Step One – Investigate the possibility of overlaying data mining algorithms directly on a data warehouse. This may require extra efforts and diplomatic skills with the IT department, but it pays off in the long run. This avoids possible errors in field names, unexpected changes in data types, and extra effort to refresh multiple data domains. The scoring can also be directly performed to live database. This is also called End-to-End data mining solution, also called Knowledge Discovery in Database (KDD). (Total time usage: 5% - 15%)

Step Two – Select a solid querying tool to build data mining files. These files closely resemble multidimensional cubes. As a matter of fact, MOLAP (multidimensional online analytical processing) serves this purpose really well. Except for APPRORI, which can use transactional data files directly (alas!), all other algorithms need "tabular" files, which are relational database files queried to produce a file with unique records with multiple fields. A number of querying tools are available for this purpose. SQL skills are highly desirable. This step can be most time consuming. (Total time usage: 30% - 75%)

Step Three – Data visualization and validation. This means both examining frequency counts as well as generating scatter plots, histograms, and other graphics, including clustering models. A graph is the best spokesperson for a correlation estimate. This step gives the researcher the first impression of what each of the data fields contains and how they may play out in the analysis. Missing data shouldn't be treated in the same manner in every situation. In certain cases, missing data is extremely diagnostic. In data mining, the outliers may be just what we are looking for, simply because they deviate from the norm. Therefore, they may hold truth in discovering previously unknown patterns. In fraud detection, it is these outliers that will flag the system to avoid loss. (Total time usage: 10% - 20%)

Step Four – Mine your data! (Total time usage: 10% - 20%)

Copyright by [illegible], 2002.

References

_____, (2001) SPSS Clementine 6.0 User's Guide. SPSS. Chicago. IL

Crowley, Bill (2000) Knowledge Management for the Information Professional. In Srikantaiah and Koenig, eds. Tacit Knowledge and Quality Assurance: Bridging the Theory-Practice Divide. Chapter 12. Informational Today Inc.: Medford, New Jersey.

Davenport T. and Prusak L. (1998) Working Knowledge. How organisations manage what they know. Harvard Business School Press. Boston, MA.

Delmater R., and Handcock M. (2001) Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence. Digital Press. Boston, MA.

Garson, G. D. (1998) Neural Networks. An introductory guide for Social Scientists. SAGE. London, UK.

Luan, J. (2002). "Data Mining and Its Applications in Higher Education" in A. Serban and J. Luan (eds.) *Knowledge Management: Building a Competitive Advantage fir Higher Education*. **New Directions for Institutional Research, No. 113**. San Francisco, CA: Jossey Bass.

Luan, J. (2000) An Exploratory Approach to Data Mining in Higher Education- A primer and a case study. Preceeding at AIR 2000. Seattle, WA.

Rubenking, N. (2001) Hidden Messages. PC Magazine. May 22, 2001

Therling, K (1995) An Overview of Data Mining at Dun and Bradstreet. DIG White Paper.

van den Eijkel, G. C. (1999) Chapter Six: Rule Induction. Intelligent Data Analysis. *Editors: Berthold, M. & Hand, D.* Springer. Milan, Italy.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").